

F_0 MODELING IN HMM-BASED SPEECH SYNTHESIS SYSTEM USING DEEP BELIEF NETWORK

Sankar Mukherjee

Yantra Software
Hyderabad
India
sankar1535@gmail.com

Shyamal Kumar Das Mandal

Centre for Educational Technology
Indian Institute of Technology Kharagpur
Kharagpur, India
sdasmandal@cet.iitkgp.ernet.in

ABSTRACT

In recent years multilayer perceptrons (MLPs) with many hidden layers Deep Neural Network (DNN) has performed surprisingly well in many speech tasks, i.e. speech recognition, speaker verification, speech synthesis etc. Although in the context of F_0 modeling these techniques has not been exploited properly. In this paper, Deep Belief Network (DBN), a class of DNN family has been employed and applied to model the F_0 contour of synthesized speech which was generated by HMM-based speech synthesis system. The experiment was done on Bengali language. Several DBN-DNN architectures ranging from four to seven hidden layers and up to 200 hidden units per hidden layer was presented and evaluated. The results were compared against clustering tree techniques popularly found in statistical parametric speech synthesis. We show that from textual inputs DBN-DNN learns a high level structure which in turn improves F_0 contour in terms of objective and subjective tests.

Index Terms— F_0 Modeling, DBN, Speech Synthesis, Bengali.

1. INTRODUCTION

Prosody plays the most important role in generating natural and intelligent speech. Prosody is a collection of supra-segmental features (duration, intonation, co-articulation pattern) which contributes additional information to speech that do not found in text. In prosody, F_0 contour has a significant contribution that is crucial to human speech perception. Knowledge of these parameters are implicit in speech signal i.e. it's hard to capture the rules governing these knowledge. Traditional shallow architectures (i.e. statistical model with few levels of computation units) are fine in limited domain but they are not capable of handling these parameters when there are lots of variations in the test set. In this context, DNN is famous for its ability to capture internal representations that become increasingly complex.

In recent years there has been a lot of interest in applying DNN in different speech processing tasks. Main advantage of

these deep learning techniques that they can learn from unlabeled data and only a limited number of labeled example are needed to fine-tune the model according to the specific tasks at hand. DNNs are basically a multilayer perceptron with many hidden layers. DBN is a class of DNN family. It is a probabilistic generative model with multiple layers of stochastic, hidden variables. Each pair of layers is treated as a Restricted Boltzmann Machine (RBM) which is a bipartite undirected graphical model with two-layer architecture. The training of DBN as described in [1] is to first initialize the weights of each layer greedily in a purely unsupervised way and then fine-tune all the weights jointly to further improve the likelihood. The resulting DBN is considered as a hierarchy of non-linear feature detectors that can learn complex statistical patterns. Rather than initialize random weights in a DNN, the weights learned by DBN can be used as the weights of a DNN. This is commonly called *pre-training* of DNN. The whole DNN can be further fine-tune by a small number of labeled training data. DBN-DNN (DNN pre-trained by DBN and fine-tune by labeled data) is successfully applied in speech, audio, image and text data [2] [3]. This advances triggered interest in applying deep learning techniques in speech synthesis tasks. In recent years, DBNs have been successfully applied to modeling speech signals, such as spectrogram coding [4], speech recognition [5], and acoustic-articulatory inversion mapping [6], where they mainly act as the pre-training methods for a deep autoencoder or a deep neural network (DNN). In statistical parametric speech synthesis domain DBNs have also been studied very recently [7] [8] [9].

Use of DBN to model F_0 contour [8] [10] was not new. But in [10] DBN was used as feature extractor for Gaussian Process Regression which is a non-parametric model. Our work was different in the sense that our prediction model was based on DNN with weight initialized by DBN. This makes the proposed model completely parametric which has many advantages like smaller foot print in contrast to non-parametric model. By doing that the proposed method could be used in hand held devices as a stand alone application. An-

other aspect was rather than discontinues F_0 contour [8] we train the DNN with continues F_0 contour which adds simplicity to the model.

This paper organized as follows. In Section 2, we will briefly review the basic techniques of RBMs and DBNs. In Section 3, we will describe the details of our proposed method. Section 4 reports our experimental results. Section 5 gives the conclusion and the discussion on our future work.

2. ARCHITECTURE OF DEEP BELIEF NETWORK

2.1. Restricted Boltzmann Machines

RBM is a special type of Markov random field that has one layer of (Bernoulli) stochastic hidden units and one layer of (Bernoulli or Gaussian) stochastic visible or observable units. There are no visible-visible or hidden-hidden connections but all the visible units are connected to all the hidden units. The weights between the connections of the visible units v and hidden units h define a probability distribution over the visible units v via an energy function [11]. Depending on the visible unit (i.e. Bernoulli or Gaussian) there are two types of energy function (i.e. Bernoulli (visible)-Bernoulli (hidden) and Gaussian (visible)-Bernoulli (hidden)) of the joint configuration (v, h) . Gaussian-Bernoulli RBMs is used to convert real-valued stochastic variables into to binary stochastic variables, which then further processed using the Bernoulli-Bernoulli RBMs. In this work Bernoulli-Bernoulli is used and its energy function is defined as

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_{i=1}^V \sum_{j=1}^H w_{ij} v_i h_j - \sum_{i=1}^V b_i v_i - \sum_{j=1}^H a_j h_j \quad (1)$$

Where $\theta = (\mathbf{w}, \mathbf{b}, \mathbf{a})$ and w_{ij} represents the symmetric interaction term between visible unit v_i and hidden unit h_j , b_i and a_j the bias terms, and V and H are the numbers of visible and hidden units. The joint distribution $p(v, h; \theta)$ over the visible units v and hidden units h , given the model parameters θ , in terms of an energy function $E(v, h; \theta)$ is defined as

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z} \quad (2)$$

where $Z = \sum_v \sum_h \exp(-E(\mathbf{v}, \mathbf{h}; \theta))$ is a normalization factor. The marginal probability that the model assigns to a visible vector \mathbf{v} is

$$p(\mathbf{v}; \theta) = \frac{\sum_h \exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z} \quad (3)$$

As there are no hidden-hidden or visible-visible connections, the conditional distributions are factorial and are given by

$$p(h_j = 1 | \mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^V w_{ij} v_i + a_j\right) \quad (4)$$

$$p(v_j = 1 | \mathbf{h}; \theta) = \sigma\left(\sum_{i=1}^H w_{ij} h_i + b_i\right) \quad (5)$$

where $\sigma(x)$ is the activation function. Here $\sigma(x) = \frac{1}{1+\exp(-x)}$ was considered. Taking the stochastic gradient descent of the negative log likelihood l the update rule for the RBM weights as

$$\Delta w_{ij}(t+1) = m \Delta w_{ij}(t) - \alpha \frac{\partial l}{\partial w_{ij}} \quad (6)$$

where α is the learning rate and m is the momentum factor used to smooth out the weight updates. General form of the derivative of the log likelihood of the data can be written as

$$-\frac{\partial l(\theta)}{\partial w_{ij}} = E_{data}(v_i, h_j) - E_{model}(v_i, h_j) \quad (7)$$

where $E_{data}(v_i, h_j)$ is the expectation observed in the training set and $E_{model}(v_i, h_j)$ is that same expectation under the distribution defined by the model. But, $E_{model}(v_i, h_j)$ is computationally very expensive to compute so the contrastive divergence (CD) algorithm [12] to the gradient is used where $E_{data}(v_i, h_j)$ is replaced by running the Gibbs sampler initialized at the data for one full step.

2.2. Deep Belief Network

A DBN is formed by Stacking a number of the RBMs learned layer by layer from bottom up. In this model, each layer captures the correlations among the activities of hidden features in the layer below. The top two layers of the DBN form an undirected bipartite graph. The lower layers form a directed graph with a top-down direction to generate the visible units. Given the training samples of the visible units, it is difficult to estimate the model parameters of a DBN directly under the maximum likelihood criterion due to the complex model structure with multiple hidden layers. Therefore, a greedy learning algorithm has been proposed and popularly applied to train the DBN in a layer-by-layer manner [1]. After learning Bernoulli-Bernoulli RBM the activation probabilities of its hidden units was treated as the data for training the Bernoulli-Bernoulli RBM one layer up. The activation probabilities of the 2^{nd} -layer Bernoulli-Bernoulli RBM are then used as the visible data input for the 3^{rd} -layer Bernoulli-Bernoulli RBM, and so on. This greedy procedure above achieves approximate maximum likelihood learning. It has been proved that this greedy learning algorithm can improve the lower bound on the log-likelihood of the training samples by adding each new hidden layer [1] [13]. AIS-based partition function estimation with approximate inference [13] used to estimate the lower bound on the log-probability.

3. PROPOSED F_0 MODELING APPROACH

As shown in Figure 1, a database of speech and corresponding text sentences was used as the training corpus. CRBLP

speech corpus [14] is employed here for the whole experiment and it consists of one male voice of age 27. STRAIGHT [15], a high-quality analysis and synthesis algorithm, was adopted to estimate the spectrum and F_0 contours with 10-ms frame rate. Continues F_0 contour was formed using step described in [16] which resulted an approximation of original F_0 contour consisting of third order polynomial segments. For the input of the neural network a set of textual features were extracted from the raw text. Table 1 illustrates the features considered for this work. The phoneme consists of 30 consonants and 16 vowels. Max syllable length 6 and max 10 syllable words was considered and which was sufficient for Bengali language. These features were re-encoded using One-of-N codes which resulted 220 binary features. All though the phoneme and syllable properties may differ for language to language. To mitigate this language dependency problem we only need to adapt language specific text analysis module. Apart from the text analysis module the whole system is completely independent of Language.

Table 1. Textual features extracted from raw text

Feature Name	No. of Features
Phoneme identity [Previous/Current/Next]	$46 * 3 = [138]$
No. of syllable in current word	[10]
Phoneme position in syllable [Forward/Backward]	$6 * 2 = [12]$
Syllable position in word [Forward/Backward]	$10 * 2 = [20]$
No. of phonemes in syllable [Previous/Current/Next]	$6 * 3 = [18]$
Vowel position in syllable [Forward]	[6]
Vowel identity in the syllable	[16]
Total	220

3.1. DBN Training

Input to the DBN was binary i.e. we used Bernoulli-Bernoulli RBM. From the corpus 7000 sentences are chosen to train the DBN. Each DBN layer was pre-trained for 50 epochs as a RBM with mini-batch of size 10. Average gradients were computed on the mini-batches and parameters were updated with a learning rate of 0.002 and a momentum of 0.95. Different architecture of DBN illustrated in Figure 3 were trained with these configuration.

3.2. DNN Training

DNN is pre-trained by the DBN which means weights of the DNN is initialized from the trained DBN. In order to fine-tune the DNN 1000 sentences (excluding the previous 7000)

are taken from the CRBLP corpus and they are phonetically aligned by HTK toolkit (5-state HMM). Finally the phonetic boundaries are manually corrected. From the 1000 sentences 500 are chosen to train, 200 to cross validation, 300 to test the DNN. Output of the DNN are log F_0 values corresponds to each phoneme state. This is because sentences are segmented using 5-state HMMs which results 5 states for each phoneme or each observation corresponds to roughly 1/5 of the phonemes. F_0 values corresponding to each state were calculated from the continues F_0 contour with the duration information generated by HTK. These F_0 values act as the output of DNN. We used mean squared error as the objective function with sparsity target of 0.05 and 0.002 weight decay. With mini-batches of 100 states backpropagation training was evaluated using a cross validation set. Loss is measured for the 10 epochs and if loss increased then learning rate was decreased by a factor of two.

The experiment was carried out on DELL precision T3600 workstation which is a 6 core computer with a CPU clock speed of 3.2 GHz, 12MB of L3 cache and 64GB DDR3 RAM. The training also used an NVIDIA Quadro 4000 general purpose graphical processing unit (GPGPU) for matrix multiplication.

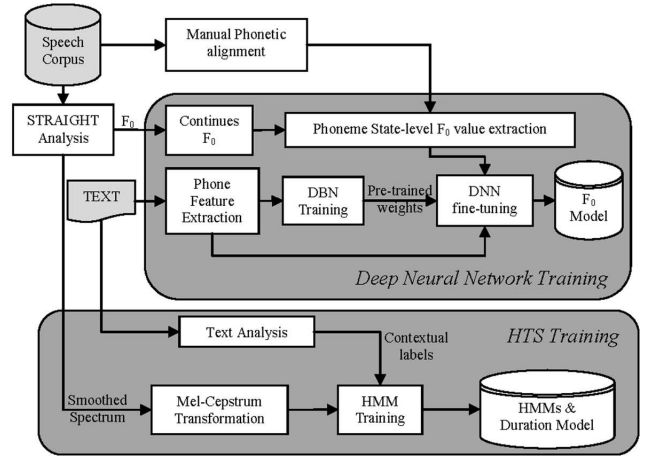


Fig. 1. Training stage of the proposed F_0 method

3.3. Synthesis Stage

Figure 2 shows the synthesis procedure of the proposed system. It has two main parts - generation of the cepstrum and duration, and prediction of the F_0 values. At the synthesis time the same features were extracted corresponding to each phoneme and they were fed to the DNN. With the previously trained weights DNN then predict the F_0 values. From the HTS phoneme state-duration model the duration of each states were extracted. Cubic spline interpolation were performed on the state F_0 values with the duration information so that the resulted F_0 has the same length. As a result, a

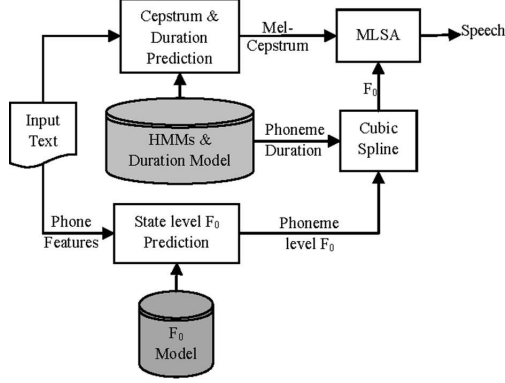


Fig. 2. Synthesis stage of the proposed method

continuous pitch contour was generated. Using MLSA filter synthesized speech was generated.

4. RESULTS AND EVALUATION

Objective evaluation are conducted to evaluate the performance of the different DBN-DNN model. The best model is chosen to further be evaluated using subjective tests. In order to compare the proposed model against the clustering tree with multi-space probability distribution (MSD-HMM) F_0 model included in the HTS synthesis engine toolkit [17], a Bengali-HTS [18] system is constructed. It is built with HTS 2.2 with 34th order MGC coefficient, 10ms Blackman window, 5ms shift and 0.53 frequency wrapping factor.

4.1. Objective Evaluation and Model Selection

For the objective evaluation two types of metrics i.e. cross-correlation (XCORR) and root mean square error (RMSE) have been calculated. Several architecture of DNN has been constructed and their performance is evaluated using these two metrics. Figure 3 illustrates the different RMSE and XCORR values of the predicted F_0 values on the test set (300 sentences). From the Figure 3 it is clear that XCORR is improving on the test set with increasing hidden layer size but the performance of RMSE is decreasing. So we choose **DBN-DNN (120U-7L)** which yields the best performance according to the two metrics (RMSE 17 and XCORR 0.64). Table 2 presents the objective test results performed on the test set between MSD-HMM and DBN-DNN. **DBN-DNN (120U-7L)** is selected for subjective evaluation.

Table 2. Result of objective test between MSD-HMM and DBN-DNN

	MSD-HMM	DBN-DNN
RMSE	25.03	17
XCORR	0.49	0.64

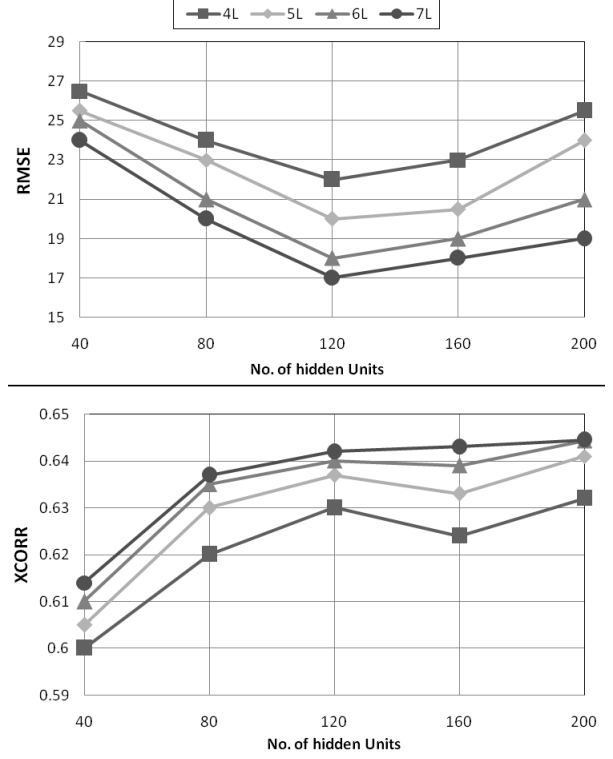


Fig. 3. RMSE and XCORR of the predicted F_0 on the test set

4.2. Subjective Evaluation

For subjective measurement, ABX is performed with 5 subjects (3 male, 2 female). All subjects are not speech experts and native speakers of Bengali. In this experiment, the 50 test sentences are synthesized using DBN-DNN (120U-7L) and MSD-HMM. Participants are asked to choose their preferred one. It can be seen from Figure 4 that 82% of the time subjects has a preference towards one of the two systems and majority (54% vs 46%) preferred proposed system and it is statistically significant with $p < 0.001$.

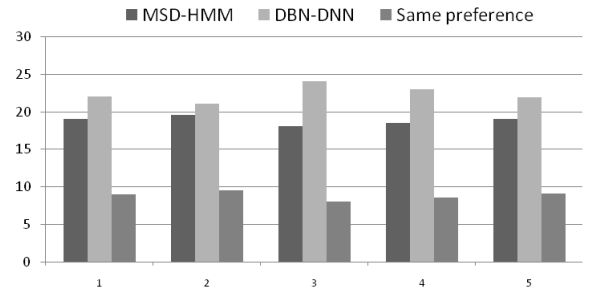


Fig. 4. ABX score of the of the two systems

5. CONCLUSION & FUTURE WORKS

In this work we have applied Deep Belief Network (DBN) to model the F_0 contour of synthesized speech which is generated by HMM-based Speech Synthesis System. DBN acted as a high level feature extractor from the raw input text. Neural network is trained for each phoneme properties i.e. for the input to the neural network textual features (phoneme identity, syllable counts etc.) and for the output normalized log F_0 values are used. Although the whole experiment is conducted on Bengali Language but it can be applied to any languages. From the objective metrics and subjective test it is found that proposed model has more preference than MSD-HMM based model which is found in many standard text-to-speech synthesis systems.

6. REFERENCES

- [1] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [2] Volodymyr Mnih and Geoffrey E Hinton, "Learning to detect roads in high-resolution aerial images," in *Computer Vision—ECCV 2010*, pp. 210–223. Springer, 2010.
- [3] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning (ICML)*. ACM, 2008, pp. 160–167.
- [4] Li Deng, Michael L Seltzer, Dong Yu, Alex Acero, Abdel-rahman Mohamed, and Geoffrey E Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *INTERSPEECH*. Citeseer, 2010, pp. 1692–1695.
- [5] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [6] Benigno Uria, Steve Renals, and Korin Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *NIPS Workshop on Deep Learning Unsupervised Feature Learning*, 2011.
- [7] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*, 2013, pp. 7962–7966.
- [8] Shiyin Kang, Xiaojun Qian, and Helen Meng, "Multi-distribution deep belief network for speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8012–8016.
- [9] Z-H Ling, Li Deng, and Dong Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.
- [10] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory, "F0 contour prediction with a deep belief network-gaussian process hybrid model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6885–6889.
- [11] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton, "Exponential family harmoniums with an application to information retrieval," in *Advances in neural information processing systems*, 2004, pp. 1481–1488.
- [12] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [13] Ruslan Salakhutdinov, *Learning deep generative models*, Ph.D. thesis, University of Toronto, 2009.
- [14] Sultana Dil Afroza Alam Firoj, Habib S. M. Murtoza and Khan Mumit, "Development of annotated bangla speech corpora, spoken language technologies for under-resourced language," in *Proceedings of (SLTU10)*. Penang, Malaysia, 2010, vol. 1, pp. 35 – 41.
- [15] Hideki Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 2, pp. 1303–1306.
- [16] Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose, and Hiroya Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, vol. 1, pp. I–509.
- [17] "HMM-based speech synthesis system (HTS)," [Online] Available: <http://hts.sp.nitech.ac.jp/>.
- [18] Sankar Mukherjee and Shyamal Kumar Das Mandal, "A bengali hmm based speech synthesis system," in *International Conference on Speech Database and Assessments (Oriental COCOSA)*, 2012, pp. 225–259.